Avoiding Spurious Local Maximizers in Mixture Modeling *

L.A. García-Escudero, A. Gordaliza,
C. Matrán and A. Mayo-Iscar
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid and IMUVA. Valladolid, Spain†

Abstract

The maximum likelihood estimation in the finite mixture of distributions setting is an ill-posed problem that is treatable, in practice, through the EM algorithm. However, the existence of spurious solutions (singularities and non-interesting local maximizers) makes difficult to find sensible mixture fits for non-expert practitioners. In this work, a constrained mixture fitting approach is presented with the aim of overcoming the troubles introduced by spurious solutions. Sound mathematical support is provided and, which is more relevant in practice, a feasible algorithm is also given. This algorithm allows for monitoring solutions in terms of the constant involved in the restrictions, which yields a natural way to discard spurious solutions and a valuable tool for data analysts.

Keywords: Mixtures; maximum likelihood; EM algorithm; constraints; eigenvalues restrictions.

1 Introduction

Finite mixtures of distribution have been extensively applied in the statistical literature to model very different types of data (see, e.g., the monographies by Titterington et al 1985, and, McLachlan and Peel 2000). This wide use has been motivated by the existence of feasible algorithms, mainly based on variations of the expectation-maximization (EM) algorithm of

*This research is partially supported by the Spanish Ministerio de Ciencia e Innovación, grant MTM2011-28657-C02-01.

†Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain.

Dempster et al. (1977). However, in practice, there are several difficulties arising from the nature of the problem, which avoid a more simple use for practitioners.

In this work, we just focus on the most extensively analyzed problem in mixture modeling, which is the problem of fitting a mixture of G normal components to a given data set $\{x_1, ..., x_n\}$ in $\mathbb{R}^p$. Moreover, we will assume that the number of components G is fixed beforehand. Our framework is that of "Maximum Likelihood (ML)" and, thus, we consider (log-)likelihoods like

$$\sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \pi_g \varphi(x_i; \mu_g, \Sigma_g) \right], \tag{1}$$

where $\varphi(\cdot; \mu, \Sigma)$ stands for the probability density function of the p-variate normal distribution with mean μ and covariance matrix Σ.

One of the main difficulties in this context is that the maximization of the log-likelihood (1) without any constraint is an ill-posed problem (Day 1969). It is well known that $\varphi(x_i; x_i, \Sigma_g)$ tends to infinity when $\det(\Sigma_j)$ approximates 0, making the target function (1) unbounded. Moreover, there exist many non-interesting local maximizers of (1), which are often referred to as spurious solutions. The choice of meaningful local maximizers (avoiding singularities and spurious solutions) is thus an important, but complex, problem.

McLachlan and Peel (2000), after showing some illustrative examples of this problem, proposed monitoring the local maximizers of (1), obtained after the application of EM type algorithms and carefully evaluating them by resorting to appropriate statistical tools. Unfortunately, this evaluation is not an easy task for practitioners without enough statistical expertise.

An alternative approach is based on considering different constrains on the Σ_g scatter matrices. Most of them are based on imposing constraints on the elements of the decomposition of the scatter matrices in the form

$$\Sigma_g = \lambda_g D_g A_g D_g'$$

(see, e.g., Banfield and Raftery 1993 and Celeux and Govaert 1995), where λ_g is the largest eigenvalue, D_g is the matrix of eigenvectors of Σ_g and A_g is a diagonal matrix. Considering the λ_g, D_g and A_j as independent sets of parameters, the idea is constrain them to be the same among the different mixture components or allow them to vary among mixture components.

Penalized maximum likelihood approaches were considered (see, e.g., Chen and Tan 2009 and Ciuperca et al 2003) to overcome the problem of unboundedness of the likelihood, and, Fraley and Raftery (2007) proposed a Bayesian regularization approach to address the problem of the spurious solutions.

Another possibility for transforming the maximization of (1) into a well-defined problem goes back to Hathaway (1985) (he also refereed to Dennis (1982), who, in turn, cited to

Beale and Thompson (oral communications)). In the univariate case, Hathaway's approach is based on the maximization of (1) under the constraint

$$\frac{\max_{g=1,\ldots,G} \sigma_g^2}{\min_{g=1,\ldots,G} \sigma_g^2} \leq c, \tag{2}$$

where $c \geq 1$ is a fixed constant and $\Sigma_g = \sigma_g^2$ are the variances of the univariate normal mixture components.

Hathaway also outlined an extension to multivariate problems based on the eigenvalues of matrices $\Sigma_j \Sigma_k^{-1}$. Unfortunately, to our knowledge, this extension has not been implemented in practical applications due to the non-existence of appropriate algorithms for carrying out the associated constrained maximization. In fact, Hathaway's attempt to provide an algorithm for this goal, even in the univariate case, addressed a different (but not equivalent problem) through the constraints

$$\frac{\sigma_{g+1}^2}{\sigma_g^2} \leq c, \text{ for } 1 \leq g \leq G-1, \text{ and } \frac{\sigma_G^2}{\sigma_1^2} \leq c.$$

These more feasible constraints were proposed as an alternative to constraints (2) in Hathaway (1983, 1986).

In this work, we consider an easy extension of constraints (2) which allows for a computationally feasible algorithm. The approach is based on controlling the maximal ratio between scatter matrices eigenvalues as it has been already considered by the authors in a (robust) clustering framework (García-Escudero et al 2008). However, our aim there was (robustly) to find clusters or groups in a data set instead of modeling it with a finite mixture. Although the two problems are clearly related, we are now using "mixture" likelihood instead of (trimmed) "classification" likelihoods. In both approaches, a constant serves to c control the strength of the constraints on the eigenvalues.

The consideration of constraints in these problems must be supported and guided by a double perspective. On the one hand, it should be soundly justified from a mathematical point of view but, on the other hand, its numerical implementation should be feasible at an affordable computational cost.

Regarding the mathematical aspects of the problem, we will prove the existence and consistency of constrained solutions under very general assumptions. Hathaway (1986) provides similar results in the univariate case that, surprisingly, have not been properly extended to multivariate cases. In any case, our results are considerably more general even in the one-dimensional setup. A direct consequence of these theoretical results is that the added constraints lead to well-defined underlying theoretical or population problems. Otherwise, the maximization of (1) through EM algorithms results in a rather "heuristic" task whose theoretical behavior would depend on the probabilistic way that the EM algorithm is initialized. In fact, it is well known that the performance of the EM algorithm in mixture modeling relies heavily on effective initializations of parameters (see, e.g., Maitra 2009).

Even though the considered constraints result in mathematically well justified problems, it is very important to develop feasible and fast enough algorithms for their practical implementation. The direct adaptation of the type of algorithm introduced in García-Escudero et al (2008) is not satisfactory at all. This type of algorithm implies solving several complex optimization problems in each iteration of the algorithm, through Dykstra's algorithm (Dykstra 1983). Instead of considering this type of algorithm, we propose adapting the algorithm in Fritz et al (2013) to this mixture fitting problem. The proposed adaptation provides an efficient algorithm for solving the constrained maximization of (1).

Gallegos and Ritter (2009a) considered other type of constraints on the Σ_j scatter matrices in (robust) clustering by resorting to the Löwner matrix ordering ($\preceq$). To be more specific, they constrained the scatter matrices to satisfy $\Sigma_j \succeq c^{-1}\Sigma_k$ for every j and k. Gallegos and Ritter (2009b) also applied this type of constraint to the mixture fitting problem. However, a specific algorithm was not given for solving those problems for a fixed value of the constant c. Instead of doing that, they proposed obtaining all local maxima of the (trimmed) likelihood and investigate the value of c needed in order that each solution fulfills one of these constraints.

Starting from constraints on the eigenvalues of matrices $\Sigma_j\Sigma_k^{-1}$ (as those originally proposed by Hathaway 1985), algorithms trying to approximate the solution of this problem were proposed in Ingrassia and Rocci (2007). In this way, they suggested algorithms based on truncating the scatter matrices eigenvalues using known lower and upper bounds on these eigenvalues. When no suitable external information is available for bounding them, they also considered a bound on the relative ratio of the eigenvalues as we do in this work. However, their algorithm for this last proposal did not directly maximize the likelihood as done in Step 2.2 of our algorithm. Their algorithm was based on obtaining iterative estimates of η, which is a lower bound on the scatter matrices eigenvalues, to properly truncate the eigenvalues and thus, as the authors commented in the paper, the proposed algorithm is quite sensitive to the choice of an initial good choice η_0 for parameter η.

Throughout this work, we are assuming that no outlying data points appear in our data set. However, the proposed methodology can be easily extended to trimmed ML approaches were a fraction α of the data is allowed to be trimmed. For instance, the way that eigenvalues ratio constraints are enforced in the proposed algorithm may be easily incorporated to the trimmed likelihood mixture fitting method in Neykov et al (2007) or to the robust improper ML estimator (RIMLE) introduced in Hennig (2004) (see, also, Coretto and Hennig 2010).

The outline of the work is as follows. We properly state the constrained problem and give mathematical properties that support its interest in Section 2. Section 3 is devoted to the description of the proposed algorithm. Section 4 presents a simple simulation study to show how the use of constraints avoids the detection of spurious solutions. In Section 5, we analyze some examples already considered in the literature. Throughout them, we

illustrate how alternative mixture fits can be explored when moving the constant c defining the constraints. Finally, we conclude in Section 6 and some hints about how to explore these alternative mixture fits.

2 Problem statement and theoretical results

Let us assume that the sample $\{x_1, ..., x_n\} \subset \mathbb{R}^p$ arises from an i.i.d random sample from an underlying distribution P. We could ideally assume that P is a mixture of G multivariate normal components but, in the presented results, only mild assumptions on the underlying distribution P will be required. Given this sample, the proposed approach is based on the maximization of the mixture log-likelihood given in (1) but with the additional:

(ER) *eigenvalues-ratio constraint*

$$M_n/m_n \leq c$$

for

$$M_n = \max_{g=1,...,G} \max_{l=1,...,p} \lambda_l(\Sigma_g) \text{ and } m_n = \min_{g=1,...,G} \min_{l=1,...,p} \lambda_l(\Sigma_g),$$

with $\lambda_l(\Sigma_g)$ being the eigenvalues when $g = 1, ..., G$ and $l = 1, ..., p$ of the Σ_g scatter matrices and $c \geq 1$ being a fixed constant.

This type of constraints simultaneously controls differences between groups and departures for sphericity. Note that the relative length of the equidensity ellipsoids axes based on $\varphi(\cdot; \mu_g, \Sigma_g)$ is forced to be smaller than $\sqrt{c}$. The smaller c, the more similarly scattered and spherical the mixture components are. For instance, these ellipsoids reduce to balls with the same radius in the most constrained $c = 1$ case.

The previously stated empirical problem admits an underlying theoretical or population counterpart:

Constrained mixture-fitting problem: Given a probability measure P, maximize:

$$E_P \left[\log \left[\sum_{g=1}^{G} \pi_g \varphi(\cdot; \mu_g, \Sigma_g) \right] \right], \tag{3}$$

in terms of the parameters $\theta = (\pi_1, ..., \pi_G, \mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G)$ corresponding to weights $\pi_g \in [0, 1]$, with $\sum_{g=1}^{G} \pi_g = 1$, location vectors $\mu_g \in \mathbb{R}^p$ and symmetric positively definite $(p \times p)$-matrices Σ_g satisfying the (ER) constraint for a fixed constant $c \geq 1$. The set of θ parameters obeying these conditions is denoted by Θ_c.

If P_n stands for the empirical measure, $P_n = (1/n) \sum_{i=1}^{n} \delta_{\{x_i\}}$, we recover the original empirical problem by replacing P by P_n.

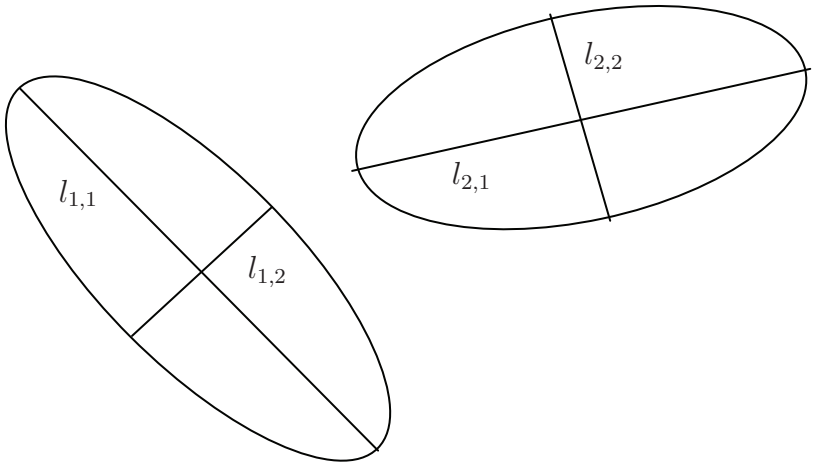

Figure 1: If $\{l_{g,l}\}$ are the length of the axes of the equidensity ellipsoids based on the $\varphi(\cdot;\mu_g,\Sigma_g)$ normal density, the constant c constraints $\max\{l_{g,l}\}/\min\{l_{g,l}\}$ to be smaller than $\sqrt{c}$.

In this section, we give results guaranteeing the existence of both empirical and population problem solutions, together with a consistency result of the empirical solutions to the population one. These two results only require mild assumptions on the underlying distribution P. Namely, we require P to have finite second moment, i.e. $E_P[\|\cdot\|^2] < \infty$, and to avoid that P is completely unappropriate for a mixture fitting approach by requesting:

(PR) The distribution P is not concentrated on G points.

This condition trivially holds for absolutely continuous distributions and for empirical measures corresponding to large enough samples drawn from an absolutely continuous distribution.

We can state the following general existence result:

Proposition 2.1 *If* (PR) *holds for distribution P and $E_P[\|\cdot\|^2] < \infty$, then there exists some $\theta \in \Theta_c$ such that the maximum of* (3) *under* (ER) *is achieved.*

The following consistency result also holds under similar assumptions:

Proposition 2.2 *Let us assume that* (PR) *holds for the distribution P with $E_P[\|\cdot\|^2] < \infty$ and θ_0 be the unique maximum of* (3) *under* (ER). *If $\theta_n \in \Theta_c$ denotes a sample version estimator based on the empirical measure P_n, then $\theta_n \to \theta_0$ almost surely.*

Recall that the original problem without the (ER) constraint is an ill-posed problem and, thus, results like the previous ones are not possible.

The proofs of these results, which will be given in the Appendix, follow similar arguments as those given for the existence and consistency results of the TCLUST method in García-Escudero et al (2008). However, a special mathematical treatment is now needed. For instance, the consistency result there needed an absolutely continuous distribution P with strictly positive density function (in the boundary of the set including the non-trimmed part of the distribution). This condition was needed due to the "trimming" approach considered by the TCLUST methodology. On the other hand, the new results for mixtures do not longer need this assumption, but they need finite second order moments to control the tails of the mixture components. The tails of the distribution were not problematic when considering trimming.

With respect to the uniqueness condition, the condition can be guaranteed when P is a mixture of G normal components once we choose a large enough c such that its scatter matrices belong to the set Θ_c. Moreover, the uniqueness condition often holds for smaller values of c. Unfortunately, stating general uniqueness results is not an easy task even in the most simple cases.

The presented approach is obviously not affine equivariant due of the type of constraints considered. Although the approach becomes closer to affine equivariance when considering large c values, it is always recommended to standardize the variables when very different measurement scales are involved.

3 A feasible algorithm

In this section, we propose an algorithm that essentially follows the same scheme adopted by standard EM algorithms in mixture fitting. However, in this new algorithm, it is very important to update the parameters in the EM algorithm in such a way that the scatter matrices satisfy the required eigenvalues ratio constraint. The proposed algorithm may be described as follows:

1. *Initialization:* The procedure is initialized `nstart` times by selecting different $\theta^{(0)} = (\pi_1^{(0)}, ..., \pi_G^{(0)}, \mu_1^{(0)}, ..., \mu_G^{(0)}, \Sigma_1^{(0)}, ..., \Sigma_G^{(0)})$. For this purpose, we propose randomly selecting $G(p+1)$ observations and computing G mean centers $\mu_g^{(0)}$ and G scatter matrices $\Sigma_g^{(0)}$ from them. The cluster scatter matrix constraints (to be described in Step 2.2) are applied to these initial $\Sigma_g^{(0)}$ scatter matrices, if needed. Weights $\pi_1^{(0)}, ..., \pi_G^{(0)}$ in the interval $(0, 1)$ and summing up to 1 are also randomly chosen.

2. *EM steps:* The following steps are alternatively executed until convergence (i.e. $\theta^{(l+1)} = \theta^{(l)}$) or until a maximum number of iterations `iter.max` is reached.

   2.1. *E-step:* We compute posterior probabilities for all the observation by using the

current $\theta^{(l)}$ as

$$\tau_g(x_i;\theta^{(l)}) = \frac{\pi_g^{(l)}\varphi(x_i;\mu_g^{(l)},\Sigma_g^{(l)})}{\sum_{g=1}^{G}\pi_g^{(l)}\varphi(x_i;\mu_g^{(l)},\Sigma_g^{(l)})}. \tag{4}$$

2.2. *M-step:* We update the $\theta^{(l)}$ parameters as

$$\pi_g^{(l+1)} = \sum_{i=1}^{n}\tau_g(x_i;\theta^{(l)})/n$$

and

$$\mu_g^{(l+1)} = \sum_{i=1}^{n}\tau_g(x_i;\theta^{(l)})x_i \Big/ \sum_{i=1}^{n}\tau_g(x_i;\theta^{(l)}).$$

Updating the scatter estimates is more difficult given that the sample covariance matrices

$$T_g = \sum_{i=1}^{n}\tau_g(x_i;\theta^{(l)})(x_i-\mu_g^{(l+1)})(x_i-\mu_g^{(l+1)})' \Big/ \sum_{i=1}^{n}\tau_g(x_i;\theta^{(l)})$$

may not satisfy the required eigenvalues ratio constraint. In this case, the singular-value decomposition of $T_g = U_g' D_g U_g$ is considered for each T_g matrix, with U_j being orthogonal matrices and $D_g = \text{diag}(d_{g1}, d_{g2}, ..., d_{gp})$ diagonal matrices. Let us define the truncated eigenvalues as

$$[d_{gl}]_m = \begin{cases} d_{gl} & \text{if } d_{gl} \in [m, cm] \\ m & \text{if } d_{gl} < m \\ cm & \text{if } d_{gl} > cm \end{cases}, \tag{5}$$

where m is some threshold value. The scatter matrices are finally updated as

$$\Sigma_g^{(l+1)} = U_g' D_g^* U_g,$$

with $D_g^* = \text{diag}\left([d_{g1}]_{m_{\text{opt}}}, [d_{g2}]_{m_{\text{opt}}}, ..., [d_{gp}]_{m_{\text{opt}}}\right)$ and m_{opt} minimizing the real valued function

$$m \mapsto \sum_{g=1}^{G}\pi_g^{(l+1)}\sum_{l=1}^{p}\left(\log\left([d_{gl}]_m\right) + \frac{d_{gl}}{[d_{gl}]_m}\right). \tag{6}$$

3. *Evaluate target function:* After applying the EM steps, the value of the target function (1) is computed. The set of parameters yielding the highest value of this target function is returned as the algorithm's final output.

Remark 3.1 *There is a closed form for obtaining m_{opt} just by evaluating $2pG+1$ times the function appearing in* (6)*. To do that, let us consider $e_1 \leq e_2 \leq ... \leq e_{2Gp}$ obtained by ordering the values*

$$d_{11}, d_{12}, ..., d_{gl}, ..., d_{Gp}, d_{11}/c, d_{12}/c, ..., d_{gl}/c, ..., d_{Gp}/c,$$

and consider any $2pG+1$ *values satisfying* $f_1 < e_1 \leq f_2 \leq e_2 \leq ... \leq f_{2Gp} \leq e_{2Gp} < f_{2Gp+1}$. *Compute*

$$m_i = \frac{\sum_{g=1}^{G} \pi_g^{(l+1)} \big(\sum_{l=1}^{p} d_{gl}(d_{gl} < f_i) + \frac{1}{c} \sum_{l=1}^{p} d_{gl}(d_{gl} > cf_i) \big)}{\sum_{g=1}^{G} \pi_g^{(l+1)} \big(\sum_{l=1}^{p} ((d_{gl} < f_i) + (d_{gl} > cf_i)) \big)},$$

for $i = 1, ..., 2Gp+1$, *and choose* m_{opt} *as the value of* m_i *which yields the minimum value of* (6).

In each M-step, the constrained maximization in (1) just needs to perform the minimization of the univariate function (6) instead of the minimization on Gp parameters in expression (3.4) given in García-Escudero et al (2008) under $Gp(Gp-1)/2$ linear constraints. This original problem was computationally expensive even for moderately high values of G or p. On the other hand, with this new algorithm, the computing times are not drastically increased with respect to other (unrestricted) EM mixture fitting algorithms.

The justification of Step 2.2 and Remark 3.1 follows exactly the same lines as in Fritz et al (2013). Once the $\tau_g(x_i; \theta^{(l)})$ weights are fixed, the maximization of the likelihood done in the M-step essentially coincides with that of the "classification" likelihood in Fritz et al (2013).

4 Simulation Study

In this section, a simple simulation study is given to see how the constrained mixture fitting algorithm actually works in practice.

The simulation study is based on a random sample drawn from a distribution P which is made of two normal components in dimension $p = 2$ with density

$$0.5 \cdot N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) + 0.5 \cdot N_2 \left(\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix} \right). \tag{7}$$

Data sets in dimensions $p = 6$ and 10 are generated by adding independent identically distributed standard normal variables to the additional coordinates.

We compare the results of the proposed mixture fitting algorithm for different values of c when $G = 2$ is assumed as known. Namely, we consider $c = 1$, $c = 6$, $c = 100$ and $c = 10^{10}$. Note that the "true" scatter matrices eigenvalues ratio for this two-component mixture is equal to 6. The value $c = 1$ yields the most constrained case when we would be searching for mixture components with scatter matrices being the same diagonal matrix and with the same value in its diagonal. $c = 100$ can be seen as a "moderate" choice of c (we do not want the length of any ellipsoid axis to be $\sqrt{100} = 10$ times larger than other) and $c = 10^{10}$ means an (almost) unrestricted case where the algorithm does not force any constraint on the eigenvalues.

In this simulation, it is needed a measure of how a mixture fit is close to another one. Given a fitted mixture $\mathcal{M}$ with parameters $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$, we define

$$z_i^{\mathcal{M}} = 1 \text{ if } \pi_1\phi(x_i; \mu_1, \Sigma_1) > \pi_2\phi(x_i; \mu_2, \Sigma_2) \text{ and 0 if it is not} \tag{8}$$

or

$$z_i^{\mathcal{M}} = \frac{\pi_1\phi(x_i; \mu_1, \Sigma_1)}{\pi_1\phi(x_i; \mu_1, \Sigma_1) + \pi_2\phi(x_i; \mu_2, \Sigma_2)}. \tag{9}$$

We can, thus, measure the "discrepancy" between two mixtures $\mathcal{M}_1$ and $\mathcal{M}_2$ as

$$\delta(\mathcal{M}_1, \mathcal{M}_2) = \min\left\{ \sum_{i=1}^{n} \left|z_i^{\mathcal{M}_1} - z_i^{\mathcal{M}_2}\right|/n, \sum_{i=1}^{n} \left|z_i^{\mathcal{M}_1} - (1 - z_i^{\mathcal{M}_2})\right|/n \right\}$$

We use the notation $\delta^{\text{Classif}}(\mathcal{M}_1, \mathcal{M}_2)$ when considering z_i as defined in (8) and the notation $\delta^{\text{Mixt}}(\mathcal{M}_1, \mathcal{M}_2)$ when considering z_i as in (9).

Let us denote by $\mathcal{M}_0$ to the mixture (7) which has generated our data set. Given that $\mathcal{M}_0$ is known, we can measure through these δ discrepancies how close a fitted mixture is to the "true" underlying mixture $\mathcal{M}_0$.

Table 1 shows the result of applying the presented algorithm with `nstart`= 1000 random initializations, as those proposed in Section 3. The performance of the algorithm for different values of the constant c is evaluated through two measurements:

(a) `Concordance`: The number of random initializations that ends up with mixtures $\mathcal{M}$ such that $\delta(\mathcal{M}, \mathcal{M}_0) < 0.2$ and $\delta(\mathcal{M}, \mathcal{M}_0) < 0.1$. That is, we are interested in the number of random initialization which lead to mixtures that essentially coincide with the "true" underlying mixture $\mathcal{M}_0$.

(b) `Spuriousness`: The number of random initializations that ends up with mixtures $\mathcal{M}$ such that $\delta(\mathcal{M}, \mathcal{M}_0) \geq 0.2$ and $\delta(\mathcal{M}, \mathcal{M}_0) \geq 0.1$ and taking strictly larger values for the target function (1) than the value obtained for the "true" solution $\mathcal{M}_0$. That is, they are spurious solutions which do not essentially coincide with the "true" underlying mixture $\mathcal{M}_0$, but with higher values of the likelihood.

To read this table, we must take into account that `Spuriousness`=0 and `Concordance`>0 are really needed for a good performance of the algorithm. With `Spuriousness`=0, we are avoiding the detection of spurious solutions that would be eventually preferred (due to the value of their likelihoods) to solutions closer to the "true" one. `Concordance`>0 gives the algorithm some chance of detecting a mixture close to the "true" solution.

We can see in Table 1 that the $n = 100$ small sample size makes easier the detection of spurious solutions. Moreover, as expected, higher dimensional cases, as $p = 10$, make easier the detection of spurious solutions too. However, small or even moderate values of c serve to avoid the detection of spurious solutions. Note that the consideration of $c = 1$ (smaller

			Concordant		Spurious	
n	p	c	$\delta < 0.2$	$\delta < 0.1$	$\delta \geq 0.2$	$\delta \geq 0.1$
100	2	1	990	990	0	0
		6	993	993	0	0
		100	657	662	0	0
		10^{10}	534	536	0	0
	6	1	989	989	0	0
		6	991	991	0	0
		100	67	83	0	0
		10^{10}	19	25	10	10
	10	1	991	991	0	0
		6	984	984	0	0
		100	3	13	0	0
		10^{10}	1	2	53	53
200	2	1	995	995	0	0
		6	989	989	0	0
		100	827	827	0	0
		10^{10}	697	698	0	0
	6	1	993	993	0	0
		6	993	993	0	0
		100	474	510	0	0
		10^{10}	236	254	0	0
	10	1	998	998	0	0
		6	998	998	0	0
		100	22	31	0	0
		10^{10}	5	7	3	4

Table 1: Number of random initializations out of 1000 (i.e., considering `nstart`= 1000 for each sample) that lead to mixtures close to the "true" one (`Concordance`) and those that lead to spurious mixtures (`Spuriousness`) with $\delta = \delta^{\text{Classif}}$.

than the true eigenvalues ratio $c = 6$) is not too detrimental. We can also see that the choice of small values of the constant c increase the chance of initializations ending up close to the "true" solution. On the contrary, the use of more unrestricted algorithms entails the detection of spurious solutions and makes harder the detection of solutions close to the true one, especially, in the higher dimensional cases (even when $n = 200$).

The results reported in this table correspond to the use of the discrepancy measure δ^{Classif} but similar results are obtained when considering δ^{Mixt}.

5 Examples

This section is based in some examples presented in McLachlan and Peel (2000) to illustrate the difficulties that spurious solutions introduces in mixture fitting problems. We see how the proposed constrained mixture fitting approach can be successfully applied to handle these difficulties.

5.1 McLachlan and Peel's "Synthetic Data Set 3"

This data set correspond to Figure 3.8 in McLachlan and Peel (2000) and it consists of 100 observations randomly generated from a heterocedastic mixture of two bivariate normal components. It was introduced there to see the high prevalence of spurious local maximizers in ML estimation of finite mixture models. Since this is a simulated data set, the "true" cluster partition is known and shown in Figure 2,(a). The associated cluster partition derived from the posterior probabilities is used to summarize the mixture fitting results.

As already commented, spurious local maximizers corresponds to solutions including populations with "little practical use or real-world interpretation". This is the case of the solution shown in Figure 2,(f). Although that solution yields a value of the likelihood higher than that corresponding to the previously presented "true" solution, it is clear that the model is fitting a small local random pattern in the data rather than a proper mixture component. The ratio between the maximum and the minimum eigenvalue is close to 1000 for this spurious solution, while it is only around 3 for the "true" one.

It is always interesting to monitor the restricted ML solutions for different values of the constant c. Figure 2 shows some of the obtained solutions for different values of constant c. In this example, we obtain cluster partitions close to the "true" solution when enforcing the relative size of the eigenvalues to be smaller than (approximately) $c = 200$. This $c = 200$ level for the constraints would imply that we are no allowing relative variabilities higher than $\sqrt{200} \simeq 14$ times in the sense of standard deviations.

We can see that no many essentially different solutions need to be examined in the proposed monitoring approach. In order to highlight this fact, we have plotted in Figure 3 the value of the constant c against the obtained scatter matrices eigenvalues ratio from the solution that the proposed algorithm returns for this value of constant c. Note that, of course, the obtained eigenvalue ratio is always smaller or equal than c but we can also see that many times the constraint is not needed to be "enforced" in the returned solution by the proposed algorithm once an upper bound on the eigenvalues ratio is posed. We say that the constraints are "enforced" by the algorithm if $M_n/m_n = c$ in (ER).

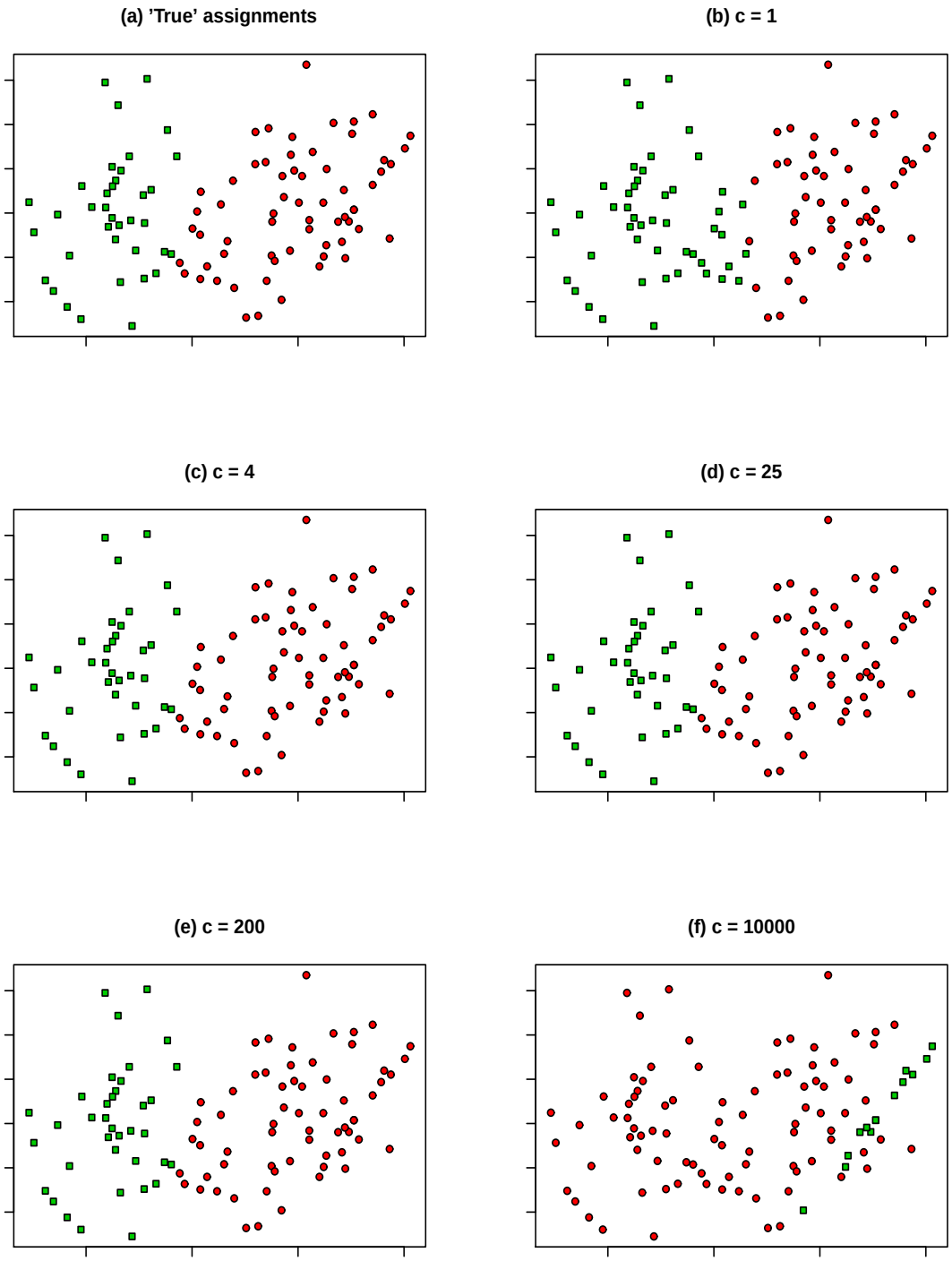

Figure 2: McLachlan and Peel's "Synthetic Data Set 3" and constrained ML clustering solutions depending on constant c.

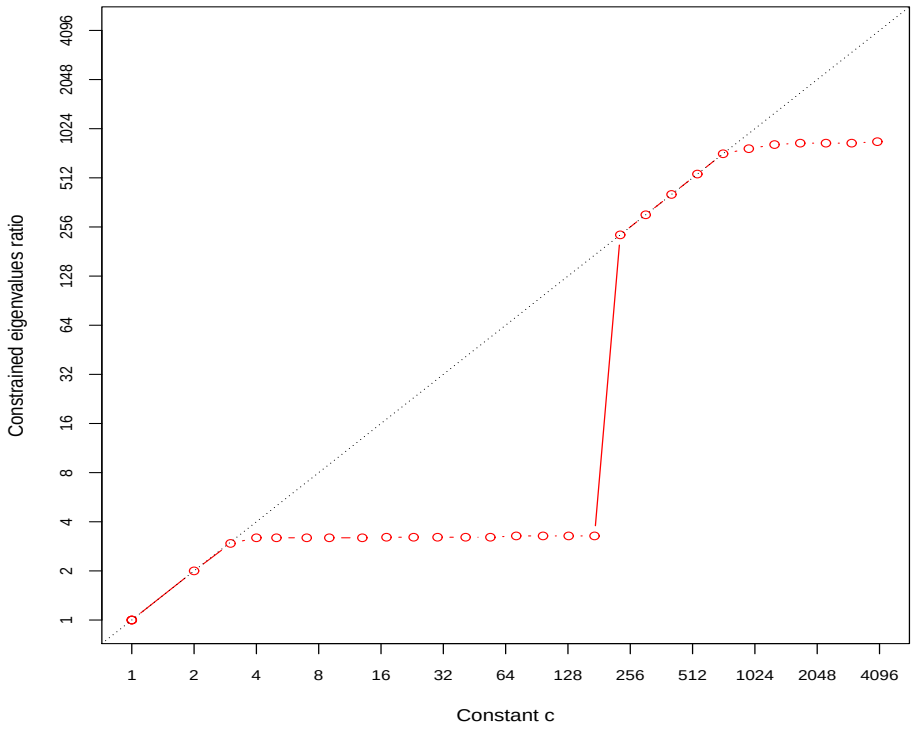

Figure 3: Plot of constant c against the "true" eigenvalues ratio for the constrained ML solution corresponding to this value of constant c (in logarithmical scales) in the McLachlan and Peel's "Synthetic Data Set 3".

5.2 "Iris Virginica" data set

The well-known Iris data set, originally collected by Anderson (1935) and first analyzed by Fisher (1936), is considered in this example. This four-dimensional ($p = 4$) data set was collected by Anderson with the aim of seeing whether there was "evidence of continuing evolution in any group of plants". Thus, it is interesting to evaluate whether "virginica" species should be split into two subspecies or not. Hence, as in McLachlan and Peel (2000)'s Section 3.11, we focus on the 50 virginica iris data and fit a mixture of $G = 2$ normal components to them.

McLachlan and Peel (2000) listed 15 possible local ML maximizers together with different quantities summarizing aspects as the separation between clusters, the size of the smallest cluster and the determinants of the scatter matrices corresponding to these solutions. After analyzing this information, an expert statistician could surely choose the so-called "S_1" solution as the most sensible solution among them, even though this solution is not the one providing the largest likelihood. The cluster partition associated to this S_1 solution is shown in Figure 4 in the two first principal components.

Unfortunately, the careful examination of such (typically big) lists of local ML maximizers is not straightforward for a non-expert users. Our proposal is to compute constrained ML

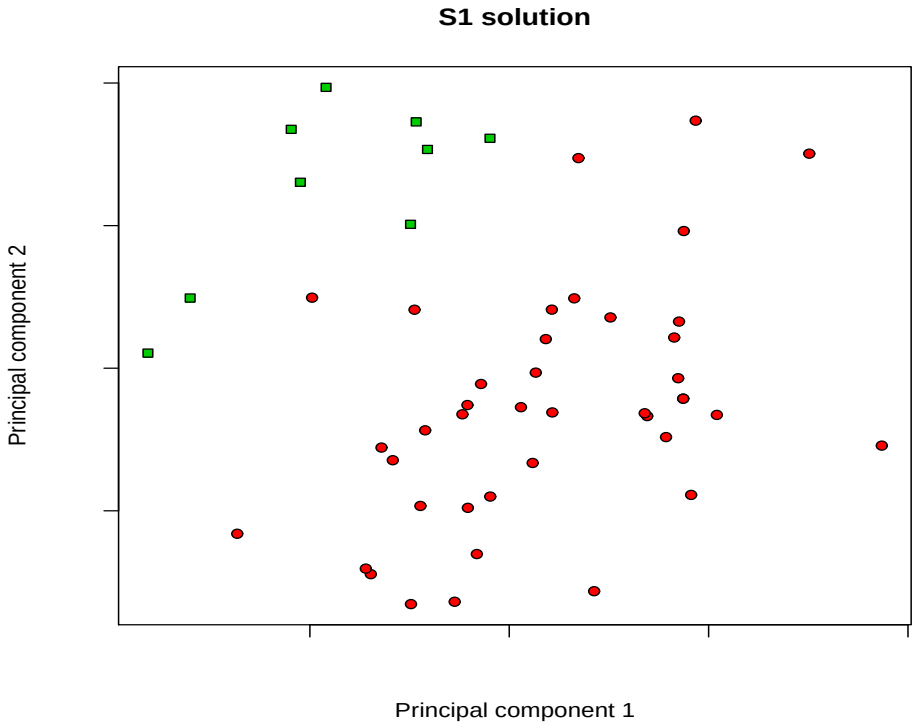

Figure 4: Plot of the first two principal components of "virginica" data set and the "S_1" solution in McLachlan and Peel (2000).

mixture fits for a grid of c values and choose a sensible one among the associated constrained solutions. This list of constrained solutions could be even more simplified by considering only those solutions which are essentially different (we will outline this further simplification in Section 6). For this example, after setting the restrictions at different c values ranging from $c = 4$ to $c = 1000$, we only get essentially the solution S_1. In fact, the differences with that solution reduce to less than one observation when analyzing the associated cluster partitions based on maximum posterior probabilities.

In order to reinforce previous claims, we show in Figure 5, a huge number of local ML maximizers obtained by running the proposed algorithm with a large $c = 10^{10}$ value. McLachlan and Peel (2000) found 51 local maxima with likelihoods greater than that corresponding to S_1 out of 1000 initializations when using the stochastic version of the EM algorithm (Celeux and Diebolt 1985). Since the proposed algorithm in Section 3 is able to visit many local maximizers, we find 787 local maximizers with higher likelihoods, than that corresponding to S_1, when considering `nstart`= 50000 and `iter.max`= 100. Thus, the number of local maxima to be explored is huge even in this very simple example. The values of the log-likelihood and the eigenvalues-ratio for these ML local maximizers are plotted in Figure 5. The same values are plotted for the constrained solutions obtained from a sequence of c values on a equispaced grid within $[1, 10^8]$ in a logarithmical scale. The corresponding values

for 13 out of the 15 solutions listed in McLachlan and Peel (2000) are also represented (2 of these solutions are not found among the obtained local maxima) and the preferred S_1 solution is also highlighted.

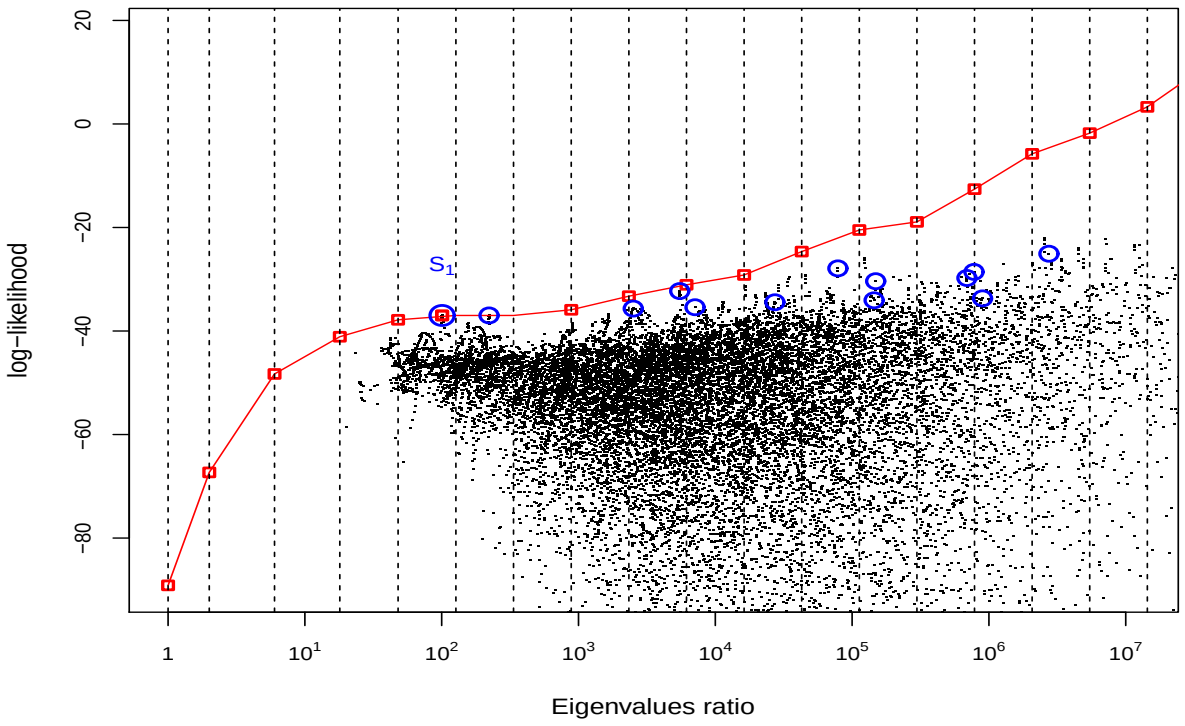

Figure 5: Log-likelihoods and eigenvalues-ratios for several local ML maximizers in the "Iris Virginica" data set and for the constrained ML solutions (□). The considered sequence of c values is represented by using vertical lines. The 13 solutions listed in McLachlan and Peel (2000), including the "S_1" solution, are enclosed by ◯ symbols.

As was previously commented, all the constrained solutions when $c \in [4, 10^3]$ are equal to solution S_1 or very close to it (with very similar log-likelihood values). There are two constrained ML solutions which have smaller eigenvalue ratios than their corresponding c values. We could see that these two solutions exactly coincide with the S_1 solution. Values $c \leq 4$ are only desirable if the user is actually interested in rather homoscedastic solutions (no axes lengths larger than 2 times the others) and $c > 1000$ leads to solutions likely to be considered as spurious since the lengths of the ellipsoid axis are very different. Thus, the practitioner, by choosing a moderate value of constant c, would have obtained the same solution (or a very close solution) as that obtained by an expert statistician.

5.3 Galaxy Data Set

This data set corresponds to the velocities (in thousand of km/sec) of 82 galaxies in the Corona Borealis region, analyzed in Roeder (1990). This data set was also used in McLachlan and Peel (2000) to show that clusters having relative very small variances (seemingly spurious solutions) may be sometimes also considered as legitimate ones. Of course, this data set has a very small sample size to answer this question and this forces us to be cautious with our statements as McLachlan and Peel did.

Figure 6 shows the galaxy data set and the solution for this data set proposed by McLachlan and Peel (2000) with $G = 6$ components. The obtained parameters for this mixture of normals can be seen in Table 3.8 of that book. In Figure 6, the two clusters suspicious of being due to a spurious local maximizer of the likelihood are labeled with letters "A" (component centered around 16.127) and "B" (centered around 26.978). Both components "A" and "B" account for 2% of the mixture distribution and their variances are 781 or 5000 times, respectively, smaller than the variance of the most scattered component.

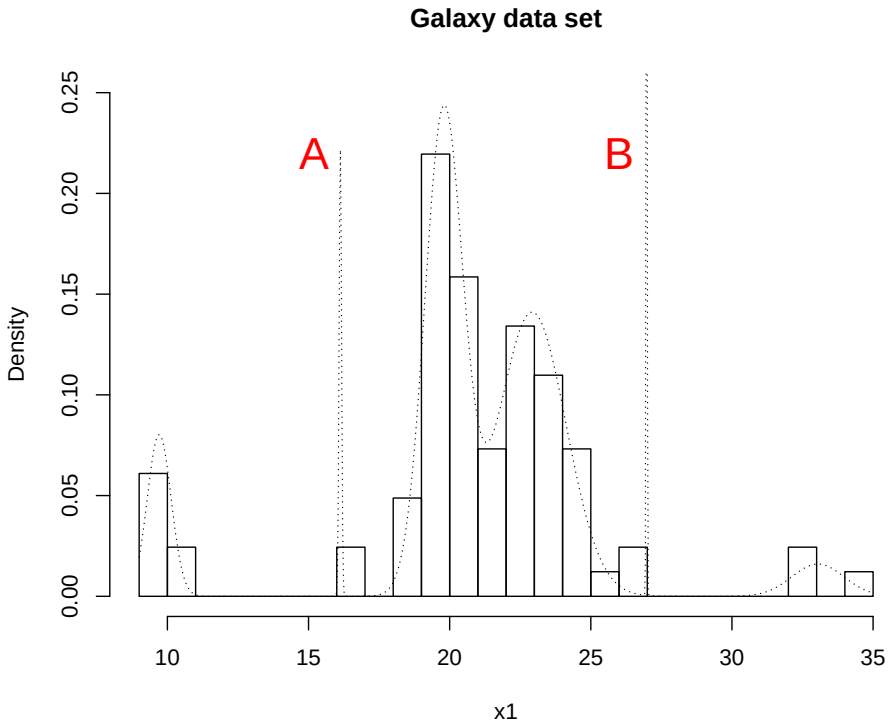

Figure 6: Histogram for the galaxy data set and the $G = 6$ normal mixture fitted in McLachlan and Peel (2000).

After examining Figure 6, it is quite logical to wonder whether components "A" and "B" can be considered just as spurious local maximizers or they are legitimate ones.

Table 2 gives the restricted ML solutions for $c = 4$, 25, 100 and 200. The "A" component is detected for this wide range of c values and, therefore, "A" can be more clearly considered

as a "legitimate" population. The "B" component is also detected when we set the c value to be greater than 100. A value $c = 100$ corresponds to allowing ten times more relative variability in the sense of standard deviation. Under the premise that this relative scatter variability was acceptable, the population "B" could be seen as a legitimate population too. McLachlan and Peel (1997) provided support for the $G = 6$ components solution and Richardson and Green (1997), following a Bayesian approach, concluded that the number of components G ranges from 5 to 7.

	$c = 4$			$c = 25$		
	π	μ	σ^2	π	μ	σ^2
	0.10	9.710	0.3306	0.13	9.710	0.1785
A:	0.04	*16.127*	0.3306	0.04	*16.127*	0.1291
	0.49	19.902	0.5655	0.33	19.765	0.4288
	0.18	22.600	0.4816	0.16	22.689	0.8068
	0.16	24.363	1.3248	0.29	23.138	3.2263
	0.04	33.044	0.8501	0.05	33.044	0.8496
	$c = 100$			$c = 200$		
	π	μ	σ^2	π	μ	σ^2
	0.07	9.710	0.1789	0.12	9.710	0.1789
A:	0.02	*16.127*	0.0166	0.02	*16.127*	0.0058
	0.30	19.703	0.3906	0.38	19.827	0.4844
	0.50	22.711	1.6615	0.37	23.013	1.1621
B:	0.01	*26.977*	0.0166	0.02	*26.978*	0.0058
	0.09	33.044	0.8501	0.09	33.044	0.8501

Table 2: Constrained solutions with $G = 6$ for the galaxy data set and different values of constant c.

6 Discussion

We have presented a constrained ML mixture modeling approach. It is based on the traditional maximization of the likelihood, but constraining the maximal ratio between the scatter matrices eigenvalues to be smaller than a fixed in advance constant c. We have seen that this approach has nice theoretical properties (existence and consistency results) and a feasible algorithm has been presented for its practical implementation.

Sometimes, the practitioner has an initial idea of the maximum allowable difference between mixture component scatters. For instance, a small c must be fixed if components with very similar scatters are expected. In any case, after standardizing the data variables,

we should always choose a small or moderate value of c just to avoid degeneracies of target function and the detection of non-interesting spurious solutions.

However, we think that more useful information can be obtained from a careful analysis of the fitted mixtures when moving parameter c in a controlled way. In fact, our experience tell us that no many essentially different solutions are needed to be examined when considering "sensible" values of c (see some examples of it in this work). This would lead to a very reduced list of candidate mixture fits to be carefully investigated. To give a more accurate picture of this idea, we propose using a grid $\{c_l\}_{l=1}^{L}$ of values for the eigenvalues ratio constraint factor c, ranging between $c_1 = 1$ and a sensible upper bound $c_L = c_{\max}$ of this ratio. For instance, an equispaced grid in logarithmical scale may be used for this grid. For this sequence of c values, we obtain the associated sequence of constrained ML fitted mixtures $\{\mathcal{M}_l\}_{l=1}^{L}$ and we can see how many "essentially" different solutions exist. In order to do that, we propose using the discrepancy measures δ^{Classif} or δ^{Mixt} introduced in Section 4 (they were introduced for the $G = 2$ case but they can be easily extended to higher number of components G). We say that two mixtures $\mathcal{M}_i$ and $\mathcal{M}_j$ are essentially the same when $\delta(\mathcal{M}_i, \mathcal{M}_j) < \varepsilon$ for a fixed tolerance factor ε, which can be easily interpreted.

Table 3 shows how many essentially different solutions can be found for the random samples used in Section 4 for the two discrepancy measures (δ^{Classif} and δ^{Mixt}) and different values of the tolerance factor ε (0.01, 0.05 and 0.1). We start from the constrained solutions obtained from 18 values of constant c taken in $[1, 10^8]$ (namely, $c = \{2^0, 2^1, ..., 2^9, 10^3, 10^4, ..., 10^{10}\}$). This table also includes the number of essentially different solutions when considering 5 random samples (instead of only one) from these mixtures enclosed in parentheses. We can see that the number of solutions is not very large, apart from the $p = 10$ and $n = 100$ cases where the sample size is not very large for the high dimension considered (in fact, a smaller number of essentially different solutions are found when considering no so large values of constant c and, thus, avoiding the detection of the most clear spurious solutions). In spite of this huge range of c values, we find 5, 4, and 3 essentially different solutions for the "Synthetic Data Set 3" in Section 5 when considering $\varepsilon = 0.01$, 0.05 and 0.1, respectively. Analogously, we have 8, 6, and 3 for the "Iris Virginica Data Set" and the same values of the tolerance factor ε. The "true" solution and the S_1 solution are always found in these smaller lists of solutions for these two examples. The discrepancy measure δ^{Classif} has been used, but similar numbers are obtained with δ^{Mixt}.

Moreover, the solutions which are not "enforced" by the algorithm (see Section 5.1) are especially interesting ones. Thus, within the list of essentially different solutions, we can obtain an even smaller list of "sensible" mixture fits by focusing only on them.

Of course, this type of monitoring approach require solving several constrained ML problems and, thus, it is only affordable if we rely on efficient and fast enough algorithms as that presented in Section 3.

		δ^{Classif}			δ^{Mixt}		
n	p	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$
100	2	1 (1)	1 (1)	1 (1)	1 (2)	1 (1)	1 (1)
	6	4 (6)	4 (6)	2 (4)	4 (6)	4 (6)	2 (4)
	10	8 (12)	7 (11)	7 (9)	8 (13)	7 (11)	7 (9)
200	2	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)
	6	1 (2)	1 (2)	1 (2)	1 (2)	1 (2)	1 (2)
	10	2 (3)	2 (3)	2 (2)	2 (3)	2 (3)	1 (2)

Table 3: Numbers of "essentially" different solutions for a grid of c values of length 18 taken in $[1, 10^{10}]$ when considering the random samples in Section 4. The numbers within parenthesis are the maximum numbers of "essentially" different solutions when considering 5 random samples from those mixtures.

A Appendix: Proofs of existence and consistency results

A.1 Existence

The proof of these results follow similar arguments as those applied in García-Escudero et al (2008) but in a mixture framework instead of a clustering one. So, let us first introduce some common notation for the clustering and mixture problems. Given $\theta = (\pi_1, ..., \pi_G, \mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G)$, let us define functions $D_g(x; \theta) = \pi_g \varphi(x; \mu_g, \Sigma_g)$ and $D(x; \theta) = \max\{D_1(x; \theta), ..., D_G(x; \theta)\}$. The mixture problem is defined through the maximization on $\theta \in \Theta_c$ of

$$L(\theta, P) := E_P \left[\log \left[\sum_{g=1}^{G} D_g(\cdot; \theta) \right] \right] \tag{10}$$

while, on the other hand, the clustering problem (classification likelihood) is defined through the maximization on $\theta \in \Theta_c$ of

$$CL(\theta, P) := E_P \left[\sum_{g=1}^{G} z_g(\cdot; \theta) \log D_g(\cdot; \theta) \right] \tag{11}$$

with $z_g(x; \theta) = I\{x : D(x; \theta) = D_g(x; \theta)\}$.

Let us consider a sequence $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1^n, ..., \pi_G^n, \mu_1^n, ..., \mu_G^n, \Sigma_1^n, ..., \Sigma_G^n)\}_{n=1}^{\infty} \subset \Theta_c$ such that

$$\lim_{n \to \infty} CL(\theta_n, P) = \sup_{\theta \in \Theta_c} CL(\theta, P) = M > -\infty \tag{12}$$

(the boundedness from below for (12) can be easily obtained just considering $\pi_1 = 1$, $\mu_1 = 0$, $\Sigma_1 = I$ and and the fact that P has finite second order moments). Since $[0, 1]^k$ is a compact

set, we can extract a subsequence from $\{\theta_n\}_{n=1}^{\infty}$ (that will be denoted like the original one) such that

$$\pi_j^n \to \pi_g \in [0,1] \text{ for } 1 \leq g \leq G, \tag{13}$$

and satisfying for some $k \in \{0, 1, ..., G\}$ (a relabelling could be needed) that

$$\mu_g^n \to \mu_g \in \mathbb{R}^p \text{ for } 0 \leq g \leq k \text{ and } \min_{g>k} \|\mu_g^n\| \to \infty. \tag{14}$$

With respect to the scatter matrices, under (ER), we can also consider a further subsequence verifying one (and only one) of these possibilities:

$$\Sigma_g^n \to \Sigma_g \text{ for } 1 \leq g \leq G, \tag{15}$$

$$M_n = \max_{g=1,...,G} \max_{l=1,...,p} \lambda_l(\Sigma_g) \to \infty, \tag{16}$$

or

$$m_n = \min_{g=1,...,G} \min_{l=1,...,p} \lambda_l(\Sigma_g) \to 0. \tag{17}$$

Lemma A.1 *Given the sequence satisfying* (12)*, if* P *satisfies* (PR) *and* $E_P[\|\cdot\|^2] < \infty$*, then only the convergence* (15) *is possible.*

Proof: The proof is the same as in Lemma A.1 in García-Escudero et al. (2008) but we need an analogous of inequality (A.8) there to be applied in this untrimmed case. This result appears in Lemma A.2 below. □

Lemma A.2 *If* P *satisfies* (PR) *then there exists a constant* $h > 0$ *such that*

$$E_P\left[\sum_{g=1}^{G} z_j(\cdot;\theta_n)\|\cdot - \mu_g^n\|^2\right] \geq h.$$

Proof: Since P is not concentrated on G points then for every $\varepsilon > 0$ there exist $G+1$ points $y_1, ..., y_{G+1}$ and $\delta = \delta(\varepsilon) > 0$ such that $P[B(y_g, \varepsilon)] > \delta$ for every $g = 1, ..., G+1$. Let us consider

$$\varepsilon_0 < \min_{1 \leq j < k \leq G+1} \|y_j - y_k\|/4$$

and $h = \delta(\varepsilon_0)\varepsilon_0^2$. We trivially have

$$E_P\left[\sum_{j=1}^{G} z_j(\cdot;\theta_n)\|\cdot - \mu_j^n\|^2\right] \geq E_P\left[\min_{g=1,...,G} \|\cdot - \mu_j^n\|^2\right] \geq h > 0.$$

□

Let us now go back to our original mixture fitting problem and let us assume again that:

$$\lim_{n\to\infty} L(\theta_n, P) = \sup_{\theta \in \Theta_c} L(\theta, P) = M > -\infty \tag{18}$$

(the bounded from below in (18) again follows from considering $\pi_1 = 1$, $\mu_1 = 0$, $\Sigma_1 = I$ and that P has finite second order moments). Displays (13), (14), (15), (16) and (17) are also considered in this mixture fitting setup.

Lemma A.3 *Given the sequence satisfying* (18)*, if P satisfies* (PR) *and that $E_P[\|\cdot\|^2] < \infty$, then only the convergence* (15) *is possible.*

Proof: The proof is trivial from Lemma A.1 just taking into account the bound

$$\begin{aligned} L(\theta, P) &= E_P\left[\log\left[\sum_{g=1}^{G} D_g(\cdot;\theta)\right]\right] \leq E_P\left[\log\left(G \max_{g=1,\ldots,G} D_g(\cdot;\theta)\right)\right] \\ &= \log G + E_P\left[\sum_{g=1}^{G} z_g(\cdot;\theta)\log D_g(\cdot;\theta)\right] = \log G + CL(\theta, P). \end{aligned} \tag{19}$$

□

Lemma A.4 *Given a sequence satisfying* (18) *and assuming condition* (PR) *and $E_P[\|\cdot\|^2] < \infty$ for P, if every π_g in* (13) *verifies $\pi_g > 0$ for $g = 1, ..., G$, then $k = G$ in* (14).

Proof: If $k = 0$, we can easily see that $L(\theta_n; P) \to -\infty$. So, let us assume that $k > 0$ and we will prove that

$$E_P\left[\log\left[\sum_{g=1}^{G} D_g(\cdot;\theta_n)\right]\right] - E_P\left[\log\left[\sum_{g=1}^{k} D_g(\cdot;\theta_n)\right]\right] \to 0. \tag{20}$$

In order to do that, we can see that

$$\begin{aligned} 0 &\leq \log\left[\sum_{g=1}^{G} D_g(x;\theta_n)\right] - \log\left[\sum_{g=1}^{k} D_g(x;\theta_n)\right] \leq \log\left(1 + \frac{\sum_{g=k+1}^{G} D_g(x;\theta_n)}{D_1(\cdot;\theta_n)}\right) \\ &\leq \log\left[1 + \sum_{g=k+1}^{G} \frac{\pi_g^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2} \exp\left(\frac{1}{2}M_n^{-1}\|x-\mu_1^n\|^2 - \frac{1}{2}m_n^{-1}\|x-\mu_g^n\|^2\right)\right] \\ &\leq \log\left[1 + \sum_{g=k+1}^{G} \frac{\pi_g^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2} \exp\left(\frac{1}{2}M_n^{-1}\|x-\mu_1^n\|^2\right.\right. \\ &\qquad\qquad \left.\left. -\frac{1}{2}m_n^{-1}\|\mu_1^n-\mu_g^n\|^2 + \frac{1}{2}m_n^{-1}\|x-\mu_1^n\|^2\right)\right] \qquad (21) \\ &\leq \log\left[1 + \exp\left(-\frac{1}{2}m_n^{-1}\min_{g=k+1,\ldots,G}\|\mu_1^n-\mu_g^n\|^2\right)\right. \qquad (22) \\ &\qquad\qquad \left. \cdot \sum_{g=k+1}^{G} \frac{\pi_g^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2} \exp\left(\frac{1}{2}m_n^{-1}\|x-\mu_1^n\|^2\right)\right]. \end{aligned}$$

Expression (21) follows from the application of the triangular inequality and (22) from the fact that $M_n^{-1} \leq m_n^{-1}$. Then, for fixed x, the expression (22) tends to 0 due to (13) and that (14) makes $\min_{g=k+1,...,G} \|\mu_1^n - \mu_g^n\|^2 \to \infty$. Moreover, (22) is uniformly dominated by a function $k_1 + k_2\|x\|^2$ by using the elementary inequality $\log(1 + a\exp(x)) \leq x + \log(1 + a)$ for $x \geq 0$ together with the assumptions (13) and (14). Since $E_P[\|\cdot\|^2] < \infty$, the Lebesgue's dominated convergence theorem finally proves (20). Note that the constraint $M_n/m_n \leq c$ has been also used for deriving the pointwise convergence and the uniform domination.

Taking into account (20) and if $\tilde{\theta}$ is the limit of the subsequence $\{(\pi_1^n, ..., \pi_k^n, \mu_1^n, ..., \mu_k^n, \Sigma_1^n, ..., \Sigma_k^n)\}_{n=1}^{\infty}$, we have $\lim_{n\to\infty} \sup L(\theta_n; P) \leq L(\tilde{\theta}; P)$. As $\sum_{j=1}^{k} \pi_j < 1$, the proof ends by showing that we can change the weights $\pi_1, ..., \pi_G$ by

$$\pi_g^* = \frac{\pi_g}{\sum_{g=1}^{k} \pi_j} \text{ for } 1 \leq g \leq k \text{ and } \pi_{k+1}^* = ... = \pi_G^* = 0, \tag{23}$$

This would lead to a contradiction with the optimality in (3) and we conclude $k = G$.

□

Proof of Proposition 2.1: Taking into account previous lemmas, the proof is exactly the same as that of Proposition 2 in García-Escudero et al. (2008). Notice that if some weight π_g is equal to 0, then we can trivially choose some μ_g and Σ_g such that $\|\mu_g\| < \infty$ and such that Σ_g satisfies the eigenvalue-ratio constraint without changing (10). □

A.2 Consistency

Given $\{x_n\}_{n=1}^{\infty}$ an i.i.d. random sample from an underlying (unknown) probability distribution P, let $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1^n, ..., \pi_k^n, \mu_1^n, ..., \mu_k^n, \Sigma_1^n, ..., \Sigma_k^n)\}_{n=1}^{\infty} \subset \Theta_c$ denote a sequence of empirical estimators obtained by solving the problem (3) for P being the sequence of empirical measures $\{P_n\}_{n=1}^{\infty}$ with the eigenvalue-ratio constraint (ER) (notice that the index n now stands for the sample size).

First we prove that there exists a compact set $K \subset \Theta_c$ such that $\theta_n \in K$ for n large enough, with probability 1.

Lemma A.5 *If P if P satisfies* (PR) *and $E_P[\|\cdot\|^2] < \infty$, then the minimum (resp. maximum) eigenvalue, m_n (resp. M_n) of the matrices Σ_g^n's can not verify $m_n \to 0$ (resp. $M_n \to \infty$).*

Proof: The proof follows similar lines as that of Lemma A.1 above by using again the bound (19). We also need a bound like in Lemma A.2 but for the empirical measure. I.e., we need a constant $h' > 0$ such that

$$E_{P_n}\left[\sum_{g=1}^{G} z_j(\cdot; \theta_n)\|\cdot - \mu_g^n\|^2\right] \geq h'.$$

This constant can be obtained by a similar reasoning as that in the proof of Lemma A.2 just taking into account that the class of the balls in $\mathbb{R}^p$ is a Glivenko-Cantelli class. □

Lemma A.6 *If* (PR) *holds for distribution P and $E_P[\|\cdot\|^2] < \infty$, then we can choose empirical centers μ_g^n's such that their norms are uniformly bounded with probability 1.*

Proof: The proof of this result follows from applying a reasoning like that in the proof of Lemma A.4. Notice that the same uniform convergence to 0 on x that was needed for proving (20) is applied here too. □

In the following lemma, we will use the same notation and terminology as in van der Vaart and Wellner (1996).

Lemma A.7 *Given a compact set $K \subset \Theta_c$, the class of functions*

$$\mathcal{H} := \left\{ \log\left(\sum_{g=1}^{G} D_g(\cdot;\theta) \right) : \theta \in K \right\} \tag{24}$$

is a Glivenko-Cantelli class when $E_P[\|\cdot\|^2] < \infty$.

Proof: Let us first consider

$$\mathcal{G} := \left\{ I_B(\cdot) \log\left(\sum_{g=1}^{G} D_g(\cdot;\theta) \right) : \theta \in K \right\},$$

where B is a fixed compact set.

We have that the class of functions

$$\{\log((2\pi)^{-p/2} \det(\Sigma)^{-1/2}) - (x-\mu)'\Sigma^{-1}(x-\mu)/2 : \mu \in \mathbb{R}^p, \Sigma \in M_{p\times p}\},$$

with $M_{p\times p}$ being positive-definite $p \times p$ matrices, is a VC class because is a finite-dimensional vector space of measurable functions. Consequently, the class $\left\{ \sum_{g=1}^{G} D_g(\cdot;\theta) \right\}$ is a VC-hull class. Applying Theorem 2.10.20 in van der Vaart and Wellner (1996) with $\phi(x) = I_B(x)\log(x)$, we obtain that $\mathcal{G}$ satisfy the uniform entropy condition. Since it is uniformly bounded, we have that $\mathcal{G}$ is a Glivenko-Cantelli class.

We can also see that there exists constants a and b such that

$$|h(x)| \leq a\|x\|^2 + b \text{ for every } h \in \mathcal{H}. \tag{25}$$

Since K is a compact set, there exist constants m and M satisfying $0 < m \leq \lambda_l(\Sigma_g) \leq M < \infty$ for $g = 1, ..., G$ and $l = 1, ..., p$. For these constants, we have

$$\begin{aligned} \sum_{g=1}^{G} D_g(x;\theta) &\geq \sum_{g=1}^{G} \pi_g (2\pi)^{-p/2} M^{-p/2} \exp(-m^{-1}\|x-\mu_g\|^2/2) \\ &\geq \exp(-m^{-1}\|x\|^2/2) \sum_{g=1}^{G} \pi_g (2\pi)^{-p/2} M^{-p/2} \exp(-m^{-1}\|\mu_g\|^2/2). \end{aligned}$$

Now take into account that $\max_{g=1,\ldots,G} \|\mu_g\| < \infty$ (recall that $\theta \in K$ with K being a compact set). Thus, we see that $\log(\sum_{g=1}^{G} D_g(x;\theta)) \geq a'\|x\|^2 + b'$. On the other hand, it is easy to see that $\log(\sum_{g=1}^{G} D_g(x;\theta)) \leq (2\pi)^{-p/2} m^{-p/2}$. Thus, a bound like (25) holds.

Finally, for every $h \in \mathcal{H}$ and B a compact set on $\mathbb{R}^p$, we have

$$\begin{aligned}\left|E_{P_n}[h(\cdot)] - E_P[h(\cdot)]\right| \;\leq\; & \left|E_{P_n}[h(\cdot)I_B(\cdot)] - E_P[h(\cdot)I_B(\cdot)]\right| \\ & + \left|E_{P_n}[(a\|\cdot\|^2 + b)I_{B^c}(\cdot)] - E_P[(a\|\cdot\|^2 + b)I_{B^c}(\cdot)]\right| \to 0.\end{aligned}$$

The result follows from the fact that $h(\cdot)I_B(\cdot) \in \mathcal{G}$ and that $E_P[\|\cdot\|^2] < \infty$. □

Proof of Proposition 2.2: Taking into account Lemma A.7, the result follows from Corollary 3.2.3 in van der Vaart and Wellner (1996). Notice that lemmas A.5 and A.6 are needed in order to guarantee the existence of a compact set K such that the sequence of empirical estimators satisfies $\{\theta_n\}_{n=1}^{\infty} \subset K$. □

References

[1] Anderson, E. (1935). "The irises of the Gaspe Peninsula", *Bulletin of the American Iris Society*, **59**, 25.

[2] Banfield, J.D. and Raftery, A.E. (1993), "Model-based Gaussian and non-Gaussian clustering," *Biometrics,* **49**, 803-821.

[3] Celeux, G. and Diebolt, J. (1985), "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Comput. Statist. Quater.*, **2**, 73-82.

[4] Chen, J. and Tan, X. (2009), "Inference for multivariate normal mixtures," *J. Multiv. Anal.*, **100**, 1367-1383.

[5] Ciuperca, G., Ridolfi, A. and Idier, J. (2003), "Penalized maximum likelihood estimator for normal mixtures," *Scand. J. Statist.,* **30**, 45-59.

[6] Coretto, P. and Hennig, C. (2010), "A simulations study to compare robust clustering methods based on mixtures," *Adv. Data. Anal. Classif.,* **4**, 111-135.

[7] Day, N.E. (1969), "Estimating the components of a mixture of two normal distributions", *Biometrika*, **56**, 463474.

[8] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Statist. Soc. Ser. B*, **39**, 138.

[9] Dennis, J.E. Jr. (1982),"Algorithms for nonlinear fitting", in *Nonlinear optimization 1981 (Proceeding of the NATO Advanced Research Institute held at Cambridge in July 1981*, (M.J.D. Powell, ed. (Cambridge, 1981), Academic Press.

[10] Dykstra, R.L. (1983), "An algorithm for restricted least squares regression," *J. Amer. Statist. Assoc.,* **78**, 837-842.

[11] Fraley, C. and Raftery, A.E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer J.,* **41**, 578-588.

[12] Fraley, C. and Raftery, A.E. (2007), "Bayesian regularization for normal mixture estimation and model-based clustering," *J. Classif.,* **24**, 155-181.

[13] Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, **7**, 179188.

[14] Fritz, H., García-Escudero, L.A. and Mayo-Iscar, A. (2013), "A fast algorithm for robust constrained clustering". *Comput. Stat. Data Anal.,* **61**, 124-136.

[15] Gallegos, M.T. and Ritter, G. (2009a), "Trimming algorithms for clustering contaminated grouped data and their robustness," *Adv. Data Anal. Classif.*, **3**, 135-167.

[16] Gallegos, M.T. and Ritter, G. (2009b), "Trimmed ML estimation of contaminated mixtures," *Sankhya (Ser. A)*, **71**, 164-220.

[17] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008), "A general trimming approach to robust cluster analysis", *Ann. Statist.*, **36**, 1324-1345.

[18] Hathaway. R.J. (1983), "Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions". Ph.D. Dissertation, Rice University, Department of Mathematical Sciences.

[19] Hathaway, R.J. (1985), "A constrained formulation of maximum likelihood estimation for normal mixture distributions," *Ann. Statist.*, **13**, 795-800.

[20] Hathaway, R.J. (1986), "A constrained EM algorithm for univariate normal mixtures," *J. Stat. Comput. Simul*, **23**, 211-230.

[21] Hennig, C. (2004), "Breakdown points for maximum likelihood estimators of location-scale mixtures," *Ann. Statist.*, **32**, 13131340.

[22] Ingrassia S. and Rocci R. (2007), "Constrained monotone EM algorithms for finite mixture of multivariate Gaussians," *Comput. Stat. Data Anal.*, **51**, 5339-5351.

[23] Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007) "Robust fitting of mixtures using the trimmed likelihood estimator," *Comput. Stat. Data Anal.*, **17**, 299308.

[24] Maitra, R. (2009) "Initializing partition-optimization algorithms," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **6**, 144157.

[25] McLachlan, G. and Peel, D. (1997), "Contribution to the discussion of paper by S. Richardson and P.J. Green," *J. Roy. Statist. Soc. Ser. B*, **59**, 772-773.

[26] McLachlan, G. and Peel, D. (2000), *Finite Mixture Models,* John Wiley Sons, Ltd., New York.

[27] Richardson, S. and Green, P.J. (1997) "On the Bayesian analysis of mixtures with an unknown number of components (with discussion)," *J. Roy. Statist. Soc. Ser. B*, **59**, 731-792.

[28] Roeder, K. (1990) "Density estimation with confidence sets exemplified by superclusters and voids in galaxies," *J. Amer. Statist. Assoc.*, **85**, 617-624.

[29] Titterington, D.M., Smith A.F. and Makov, U.E. (1985), *Statistical analysis of finite mixture distributions*, Wiley, New York.

[30] Van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*, Wiley, New York.